Spring 2016

Investigating the Relationship Between Latent Driving Patterns and Traffic Safety Using Smartphone-Based Mobile Sensor Data

Kenneth Wynne
Old Dominion University

INVESTIGATING THE RELATIONSHIP BETWEEN LATENT DRIVING PATTERNS AND TRAFFIC SAFETY USING SMARTPHONE-BASED MOBILE SENSOR DATA

by

Kenneth Wynne
B.S. May 2013, Virginia Polytechnic Institute and State University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of
Requirements for the Degree of

MASTER OF SCIENCE

CIVIL ENGINEERING

OLD DOMINION UNIVERSITY
May 2016

Approved by:

Rajesh Paleti (Director)

Mecit Cetin (Member)

Hong Yang (Member)

ABSTRACT

INVESTIGATING THE RELATIONSHIP BETWEEN LATENT DRIVING PATTERNS AND TRAFFIC SAFETY USING SMARTPHONE-BASED MOBILE SENSOR DATA

Kenneth Wynne
Old Dominion University, 2016
Director: Dr. Rajesh Paleti

Crash frequency modelling has been used in the past as an attempt to quantify the expected number of crashes occurring on a certain segment of roadway given a set of variables and factors describing the roadway segment and the traffic along that segment. These models are referred to as the Safety Performance Functions (SPFs) in the Highway Safety Manual (HSM). In past studies, these SPFs have focused primarily on roadway geometric information along with limited traffic exposure data such as traffic volume. Alternate data sources for probe vehicle data are increasingly available and this research sought to exploit this new information in order to obtain an improved model. Specifically, this research aims to make use of the accelerometer sensors in smartphones to extract microscopic traffic measures that can serve as better indicators of driving patterns. The study focused on crash frequency along roadway segments in the Hampton Roads region. To start-off, mobile sensor data was collected by driving along major roadways in the Hampton Roads region during the evening peak period (4 to 6 pm). Next, this data was overlaid on the transportation network to map probe data and the roadway segments. Then, several acceleration and deceleration metrics were calculated for each roadway using the mobile sensor data. Subsequently, these metrics were appended to the VDOT crash data for the past one year. Supplementary data sources were used to assemble information regarding roadway inventory data and traffic exposure information. Next, statistical model estimation was

undertaken to identify the factors affecting crash frequency along major interstates in Hampton Roads.

The results indicate that when comparing a model based solely on roadway geometrics to a model including both roadway geometrics and probe vehicle data, the combined model was a significant improvement. Several probe vehicle data parameters capturing microscopic traffic conditions were significant in the final model. Lastly, elasticity analysis was undertaken to quantify the relative impact of different factors in the model. With regard to statistical modeling, this research considered both a Poisson and a negative binomial model that served as standard models for crash frequency modeling in the literature. The negative binomial model was found to be a significant improvement over the Poisson model. Previous research has indicated that negative binomial models tend to perform better than Poisson models when there is over-dispersion present in the dataset. This research supports this claim. Overall, this research has determined that the addition of probe vehicle data to roadway inventory data and the usage of a negative binomial model have proved to provide a robust crash frequency model.

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to my advisor, Dr. Rajesh Paleti, for his support, encouragement, and guidance throughout the development of this thesis. I would also like to thank Dr. Mecit Cetin and Dr. Hong Yang for both serving on my thesis committee and providing their expert knowledge and insight towards this research. In a similar vein, I would like to thank Olcay Şahin for his mentorship, guidance, and work he contributed towards the research.

I would like to thank my family and friends for their continued support not only through the development of this thesis but, for my entire academic and professional careers. Without their encouragement and advice I would not be where I am today.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

In the United States there were 32,675 fatalities as a result of motor vehicle crashes in 2014 and current trends show that an increase of about 8.1 percent is expected in 2015 (NHTSA, 2015). In the year 2014, in Virginia alone, 700 people were killed and 63,384 people were injured in a total of 120,282 motor vehicle accidents (DMV, 2014). The Federal Highway Administration (FHWA) is anticipating a compound annual growth rate in vehicle miles travelled (VMT) of approximately 1.04 percent through the year 2033 (FHWA, 2015). In spite of recent vehicle safety improvements, alongside of improved roadway design practices, roadway crashes remain a serious issue especially considering the anticipated increase in VMT. These crashes not only cause injury and loss of life, but they also cost a considerable amount to the people involved. For instance, in 2010, the economic costs of motor vehicle crashes in the United States totaled $242 billion. These costs come from not only from the damage to vehicles and the medical bills of the injured but, also include items such as $28 billion due to congestion (Blincoe *et al.*, 2015).

When the spatial distribution of crashes over any transportation network is analyzed, it is common to observe hotspots (e.g., major merge areas, bottlenecks) where the crash risk is relatively high. These statistics suggest that while certain hotspots may be unsafe primarily due to the geometric features of these locations, in many cases the safety risk seems to be an outcome of the unsafe driving patterns (e.g., sudden lane changes) along the roadway stretching downstream and/or upstream of the actual crash locations. Even though there is plenty of research on correlating safety measures to roadway characteristics and some elements of traffic flow (e.g., AADT, average speed), there is no significant literature on analyzing the correlation between high-resolution dynamic speed and/or acceleration data and crash risks along highway

segments. It is important to determine the direct vehicular behavior alongside of the roadway characteristics associated with higher crash rates in order to make progress towards reducing the total number of crashes.

Speeding has long been seen as an unsafe driving behavior and historical crash records support this claim. In the year 2012, 30% of crash fatalities involved a speeding vehicle (NHTSA, 2014). It is within reason to assume that it is possible to predict other driving characteristics which contribute to crashes as well. Collecting such high-resolution data is now feasible with the mobile consumer devices such as smartphones and on board diagnostic (OBD) devices. Smartphones are now equipped with sensors capable of recording vehicle performance data at a very fine temporal resolution (Zhen and Qiang 2014). These sensors can provide a rich dataset, i.e., high resolution speed and acceleration profiles, that can be used for identifying unsafe driving patterns. In fact, several auto insurance firms (e.g., Progressive's Snapshot) have been experimenting with monitoring driving activity (e.g., hard-brakes per mile) through OBD devices to assess and valuate the crash risk of individual drivers. However, there is no significant research on investigating the potential use of high-resolution data from mobile sensors or smartphones in understanding crash risks and safety measures for highway sections.

Some research has been conducted to predict what type of vehicular behavior and roadway characteristics lead to crashes with varying levels of success. Different approaches have been taken in regards to data collection, model selection, and model implementation (Mannering and Bhat, 2014). In this context, the objective of the current study is to identify unsafe driving patterns using mobile sensor data and explore the relationship between these latent driving patterns and traffic crash incidences. This research will take a simplistic and direct approach to data collection by using these new technologies to obtain data directly from probe vehicles

themselves. Unlike most previous research, this information will allow for the most relevant and rich data to be made available for statistical analysis. The goal of this approach is to obtain a more accurate model which can determine crash frequency for a wide variety of roadway segments. This research will seek to achieve the following goals.

1. Collect a rich and robust dataset of driving behavior from probe vehicles in Hampton Roads, Virginia.
2. Develop a statistical model that can accurately predict the number of expected crashes in a year for each segment of interstate roadway in Hampton Roads, Virginia.
3. Develop parameter estimates for this model using the Virginia Department of Transportation's crash database for the region.
4. Demonstrate the improvement in model fit due to high resolution data from smartphone mobile sensors.

CHAPTER 2

REVIEW OF LITERATURE

Crashes are rare and random events. So, the number of observed crashes at any given location can fluctuate year-to-year even if all the observable crash causation conditions remain the same between the two years. If the observed crash frequency is very high in one year, then it is more likely to be followed by relatively lower crash frequency in the next year, and vice-versa. This effect is referred to as the 'Regression-To-Mean Bias'. This inherent variation in observed crash frequency poses a challenge to evaluating the effectiveness of different safety countermeasures. For instance, it is unclear if the reduction (or increase) in crash occurrences is due to random fluctuation or the safety countermeasure. To address this problem, safety analysts rely on estimates of the long term average crash frequency, also referred to as 'Expected Crash Frequency', as a proxy for crash risk. The observed crash frequency across several locations is used to statistically estimate the expected crash frequency. Expected crash frequency modelling is a reliable method for determining the safety of a segment of roadway. This technique seeks to determine the long term average number of crashes per a given unit of time by developing a correlation between certain explanatory variables and the number of observed crashes relating to them. The intricacies of these models lie in the determination of which explanatory variables to consider and the type of model employed.

2.1 Explanatory Variables

Choosing which explanatory variables to consider is an important aspect of modelling crash frequency. Previous studies have looked at explanatory variables primarily in two categories, physical characteristics of the roadway and data collected regarding vehicles travelling along the roadway of study. According to Ogle (2005), the majority of these early

studies focused on physical characteristics of the roadway due to a lack of consistent and accurate data collection means. New studies have been conducted as vehicular data collection has become more accessible due to vehicles having cheap on board sensors.

Eustace, Aylo, and Mergia (2015) conducted research which focused solely on the physical conditions of the roadway and the driver's age. This type of data is was easy to collect due to it already being included with the crash data they had obtained for their study. Each of their explanatory variables, other than traffic volume counts, was categorical in nature. Considering the nature of the data source, their research could come to limited conclusions. Research conducted by Shankar *et al.* (1997) considered roadway geometry for the majority of explanatory variables when modelling accident frequencies but, this research also included factors such as annual average daily traffic and truck volumes. This information is more detailed than simply looking at roadway geometry because it begins to explore the road users themselves. Unfortunately, data such as this is unable to capture the actual flow and movements of individual vehicles. It is difficult to develop an accurate representation of expected crash frequencies when the characteristics of the actual vehicles travelling the corridor are not considered. Moreover, according to research by Mekker *et al.* (2015), the overall congested crash rate in the state of Indiana is 24.1 times greater than the uncongested crash rate. According to this finding it is important to be able to capture data in a congested roadway state. Simple aggregate measures such as average daily traffic and truck volumes cannot capture these differences between congested and uncongested conditions. Probe vehicle data, on the other hand, can be used to capture the acceleration and deceleration profiles that serve as reliable indictors of congested traffic conditions. Naturalistic driving behavior is data that is collected regarding the movements of the actual vehicle itself through space and time. Previous studies have relied on simulation

models to capture this type of data. Gettman and Head (2003) used microscopic simulations to develop surrogate safety measures in order to model crash frequency. This method of data collection allows the researcher to control for every aspect of the simulation while being able to alter the simulation to fit different scenarios. Multiple simulation inputs may be evaluated in a short period of time to get the most accurate results. A limitation in simulation based models is the fact that real world, real time observations are not directly accounted for.

Recent studies have focused on obtaining and using data collected directly in the field to develop more accurate crash frequency models (Mannering and Bhat, 2014). Onboard diagnostics (OBD) systems were originally developed to reduce vehicle emissions but, are now regularly used in transportation research to obtain the aforementioned naturalistic driving behavior data (Jun, 2006). Ogle (2005) focused primarily on data obtained using an OBD which collected vehicle travel data from the vehicle's on board computer and The Global Positioning System (GPS) which collected data through a satellite receiver in the vehicle. These were relatively new tools which could collect data directly from the vehicle instead of relying solely on outside sensors. Ogle (2005) determined that GPS is a reliable tool for measuring speed given an adequate number of connected satellites but, GPS can be unreliable in areas of bad weather or overhead obstacles such as in tunnels.

Another option when considering probe vehicle data is using data that is crowd-sourced, collected, and combined into a dataset by a third party source. Mekker *et al.* (2015) relied on crowd-sourced data for their research on determining crash rates based on traffic congestion. This data source has the benefit of allowing the researchers to have a more robust dataset that encompasses a greater length of time. The data can be collected and stored for multiple years rather than only being available for the duration of research period. This allows the researcher to

have access to probe vehicle data that was collected around the time that actual accidents occurred. A negative aspect of this source of data lies within the fact that all data is collected in an uncontrolled manner. This may cause some bias in the dataset if an overly passive or overly aggressive driver has collected the majority of the data that the researcher is using.

Wåhlberg (2004) looked at the acceleration profiles of busses as a potential indicator of crash frequency. This study concluded that driver acceleration behavior could be used as a predictor of accidents but, due to some discrepancies between samples it is difficult to determine the validity of this finding. For this study the acceleration data was recorded on-board using a g-analyst which measured the acceleration at 10 Hertz to 100th of 1g (9.81 m/s2) accuracy. This tool did not measure the acceleration from the vehicle directly but, simply measured the g-force felt by the bus starting and stopping. This may have resulted in errors due to the vehicle not producing the data itself.

A potential source for speed data could be crash reports that were completed at the scene of an accident by the police. This would appear to be a simple way to obtain a piece of driving behavior but, according to Shinar *et al.* (1983) speed should not be obtained from a police crash report. This research concluded that the police may be under a lot of stress during incident investigations and may not be able to accurately determine the speed at which the driver was going. Also, the driver himself may underreport the estimated speed which they were travelling in an attempt to lessen the likelihood of receiving additional infractions for an incident. Alternatively, speed limit may serve as a better proxy for traffic speed.

Due to the insight that the roadway attributes and traffic characteristics provide for crash frequency modelling, it is important to consider both of these types of data simultaneously. Many recent studies have done this to create a more comprehensive model.

2.2 Crash Data

The previously mentioned explanatory variables are considered to attempt to predict the occurrence of a vehicular crash. The actual crash data the model uses is an important aspect of a successful crash frequency model. The majority of previous works relied on police crash reports for crash data. Using this data allows the researcher to assign a specific crash location to each incident recorded (Mannering and Bhat, 2014). After geo-coding crash locations, all crash occurrences within a certain geographical boundary (eg: roadway segment, intersection, or county) over a one year time period are aggregated to obtain to the observed yearly crash frequency. In cases when the time period is different from one year, the effective yearly crash frequency rate is calculated by dividing the aggregated crash frequency with the number of years in the time period.

Unfortunately, this data is not always accurate and, more importantly, not all crash data is reported in the first place. Literature suggests that underreporting in crash data may result in significant bias if this phenomenon is not considered in the model. Previous research has indicated that underreporting is most likely to occur in incidents where little to no damage occurs (Yamamoto *et al.*, 2008). Amoros *et al.* (2006) conducted research which attempted to measure the amount of underreporting in crashes in France by comparing the reported crashes to the Rhône road trauma registry. This registry contains information regarding all road crashes within the Rhône County where the occupants sought medical attention. The study concluded that according to its research the police reporting rate within this county is only 37.7%. Research by Kim *et al.* (1995) reflected this idea that crashes where minor to no injury occurred are sometimes not reported by the police. This measurement of underreporting should be considered

when conducting any sort of crash frequency analysis. Not all crashes may be accounted for and this will be a source of error in the model.

A benefit of using crash databases that were created using police reports is the amount of data that is included in the report itself. Mekker *et al.* (2015) used crash reports provided through a state crash database. The crash data provided included specific information such as the number of vehicles involved, number of trailers involved, and whether or not a construction zone was associated with the crash. This information was able to be included in the model due to the detail of the crash database itself. If this information were not present then the study would not be able to include such factors in the analysis.

2.3 Crash Frequency Modelling Techniques

Crash-frequency data is count in nature, *i.e.,* observed crash frequency is a non-negative integer number without a pre-specified upper limit (*i.e.,* it is not bounded from above). So, simple regression techniques that deal with continuous data are not suited for modeling crash frequency. The Poisson and Negative Binomial models have served the standard workhorse models for modeling count data. In transportation safety, count models are used to develop Safety Performance Functions (SPFs) that quantify the frequency of crash occurrences at any given location or region (Qin *et al.,* 2005; Ahmed *et al.,* 2011; Narayanamoorthy *et al.,* 2013). The Poisson model makes the restrictive assumption that the mean and variance in the count data are the same – referred to as the '*equi-dispersion*' property. So, Poisson model cannot handle situations where the mean is less than variance (*under-dispersion*) or mean is greater than variance (*over-dispersion*). In fact, crash-frequency data typically exhibits over-dispersion. Under-dispersion is rarer than over-dispersion but, it is sometimes present in crash-frequency data. Typically, under-dispersion in crash-frequency data is observed when the sample mean

value is low and the sample size is small (Lord and Mannering, 2010; Mannering and Bhat, 2014; Hinde and Demétrio, 1998; Lord, Geedipally, and Guikema, 2010; Li *et al.*, 2013).

Some research has turned to a negative binomial distribution model after noting the limitations of Poisson regression models. Specifically, the negative binomial distribution is better suited to handle over-dispersed datasets. For instance, Eustace, Aylo, and Mergia (2015) assumed a negative binomial distribution for their generalized linear model when attempting to predict crashes that were occurring at merging and diverging areas on an interstate freeway. This method was chosen primarily as a means to account for over-dispersion in the dataset. The negative binomial model l is by far the most commonly used model in crash frequency modeling. However, the limitation of the negative binomial distribution is the fact that it cannot handle under-dispersed data which tends to occur when there are small sample sizes or low sample mean values (Lord and Mannering 2010).

Research conducted by Miaou (1994) concluded that it is important to keep in mind both the Poisson and negative binomial modelling techniques when develop crash frequency models. This research indicated that the Poisson model should first be considered as an initial model to develop the relationship between the explanatory variables and crashes. If there is high over-dispersion present then negative binomial or other more advanced models should be considered.

2.4 Summary

Researchers have taken various approaches to crash frequency modelling over the years. Recent advances in technology have allowed researchers to collect more accurate data than ever before. Data is now able to be obtained through sensors that reside within the vehicles themselves. Based on this literature review it is important to consider vehicle trajectories instead of just roadway characteristics when modelling crash frequency. Researchers have found that

using emerging data sources such as naturalistic driving behavior may provide a more accurate model (Mannering and Bhat 2014). This research will seek to build upon previous research by employing a negative binomial model using OBD and GPS data collected from probe vehicles as well as roadway geometric characteristics and incident data to develop a comprehensive crash frequency model.

CHAPTER 3

METHODOLOGY

3.1 Poisson Regression Model

The most simplistic starting point for crash frequency modelling lies in the Poisson regression model. Typical regression modelling starts with a least-squares model but, this type of model cannot be applied to crash frequency models due to the dataset being count in nature. Crashes are non-negative integer values and are therefore not continuous values. A least-squares regression requires the data to be continuous and is not applicable to crash data. Let *s* be the index for the roadway segment (*s* = 1,2 …, *S*) and *S* be the total number of roadway segments in the study area. When applied to crashes along a roadway segment the probability of observing a number of crashes *y* in one year is given by:

$$P(Y = y_s) = \frac{e^{-\lambda_s}(\lambda_s)^{y_s}}{y_s!} \qquad \text{Equation (1)}$$

Where P(y_s) is equal to the expected number of crashes in a year and λ is the Poisson parameter for that specific roadway segment. Poisson models are parameterized by specifying a value for λ as a function of a set of explanatory variables. This allows the model to predict the number of crashes based on that set of explanatory variables. The probability mass function (PMF) for a Poisson distribution can be seen in Figure 1. As can be observed, as lambda increases, the count outcome associated with peak probability continues to shift to the right.

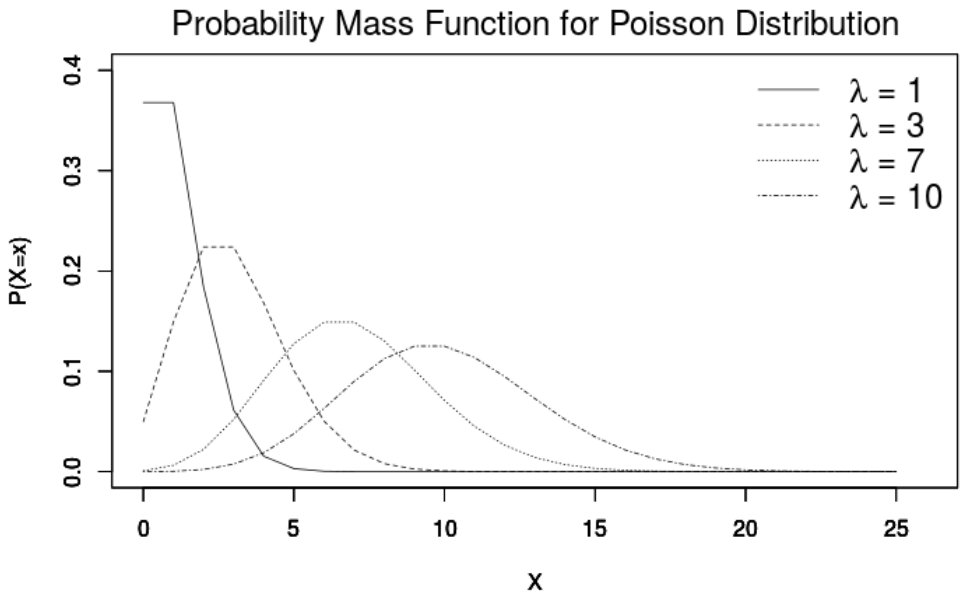

Figure 1. Probability Mass Function for Poisson Distribution

When using a Poisson model it is assumed that the expected value and the variance both equal λ. If this is not true then a Poisson model is not the best fit model for the dataset. Researchers have also discovered that the Poisson model does not fit well with data that exhibits over- or under-dispersion (Lord and Mannering 2010). It is not uncommon for crash data to represent these characteristics and therefore, other models should be considered.

3.2 Geometric Model

According to the Geometric model, the probability of observing y_s crashes conditional on the probability parameter p_s is given by:

$$P(Y = y_s) = {p_s}^{y_s}(1 - p_s); p_s \in [0,1] \quad \text{Equation (2)}$$

The expected value and variance of geometric distribution are given by:

$$E(Y) = p_s(1 - p_s)^{-1} \; and \; Var(Y) = p_s(1 - p_s)^{-2} \quad \text{Equation (3)}$$

3.3 Negative Binomial Regression Model

Researchers have turned to the negative binomial regression model as a solution to the Poisson regression's limitations. This model has an added dispersion parameter, r_s, which is an assigned value greater than zero. The probability of observing y_s crashes in a given year is given year conditional on the expected value parameter λ_s is given by:

$$P(Y = y_s) = \left(\frac{r_s}{r_s+\lambda_s}\right)^{r_s} \frac{\Gamma(r_s+y_s)}{\Gamma(y_s+1)\Gamma(r_s)} \left(\frac{\lambda_s}{r_s+\lambda_s}\right)^{y_s} \quad \text{Equation (4)}$$

Where Γ is the gamma function defined as follows:

$$\Gamma(t) = \begin{cases} \int_{x=0}^{\infty} x^{t-1}e^{-x}dx \text{ for positive non} - \text{integer } t \\ (t-1)! \text{ for positive integer } t \end{cases} \quad \text{Equation (5)}$$

The variance of the negative binomial distribution is $\lambda_s + \frac{{\lambda_s}^2}{r_s}$ which is always greater than the expected value parameter λ_s. This condition allows the negative binomial distribution to be well suited for over-dispersion; when considering crash data the variance is often higher than the mean. Figure 2 shows how changes in lambda effects the probability mass function while Figure 3 shows how changes in the over-dispersion parameter, r, effects the probability mass function.

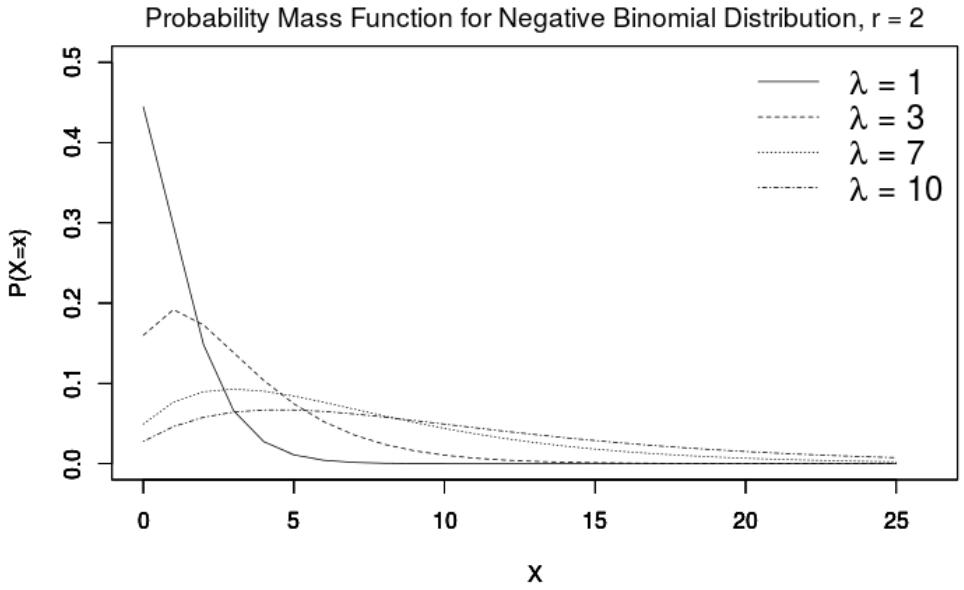

Figure 2. Probability Mass Function for Negative Binomial Distribution – r Constant

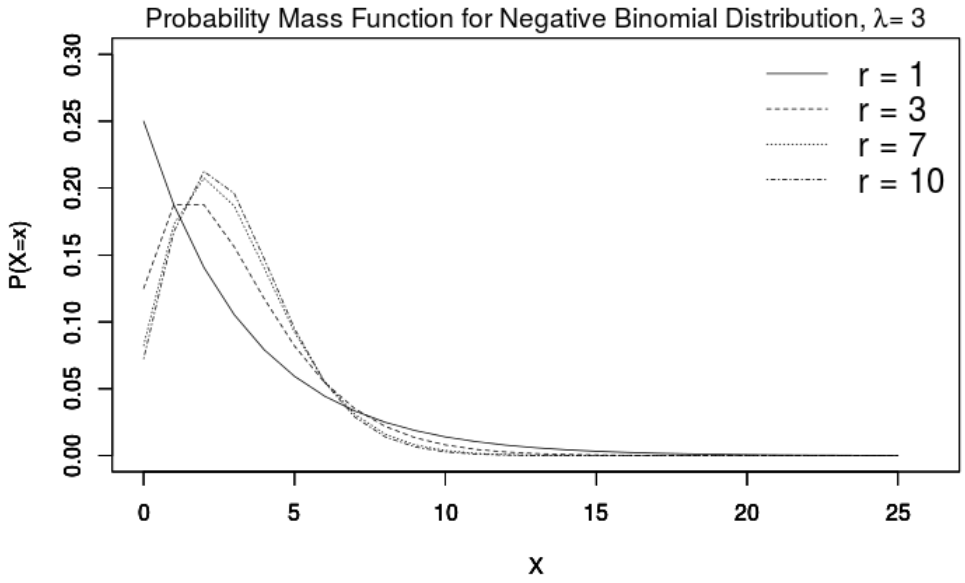

Figure 3. Probability Mass Function for Negative Binomial Distribution – Lambda Constant

As r approaches large values the negative binomial model reverts back into the Poisson model. This effect can be observed in Figure 4.

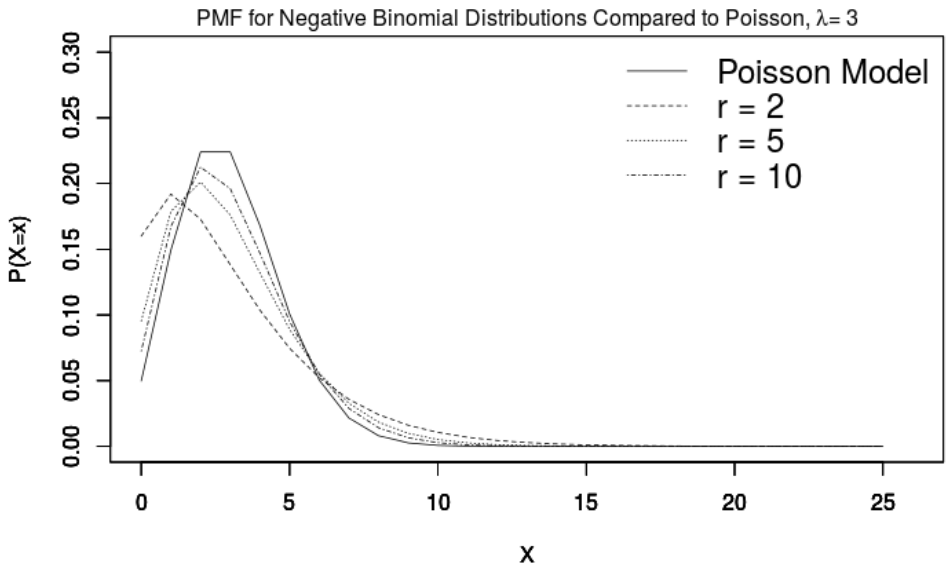

Figure 4. PMF for Negative Binomial Distributions Compared to Poisson

A limitation of the negative binomial model is when the count data is considered under-dispersed because the negative binomial model assumes an over-dispersed dataset.

CHAPTER 4

DATA ANALYSIS

4.1 Crash Database

Vehicle crash data was obtained through a Virginia Department of Transportation (VDOT) database that includes all reported crashes from October 2014 to October 2015 for the entire Hampton Roads Region. There were many records in the database which were affiliated with disabled vehicles. These records were omitted because the study is only interested in actual vehicle crashes. This raw data contained 111 characteristics for each crash. Some of this information is administrative in nature such as who recorded the crash, how it was recorded, and who last modified the report; these variables were not beneficial in the analysis. The database also recorded the type of crash (vehicle accident, multi-vehicle accident, or tractor trailer accident) and time impact severity of the crash (< 30 min., 30 min. to 2 hours, or > 2 hours). However, this study considered only total crash frequency instead of crash frequency by type and severity. So, these variables were not used in the analysis. One variable of particular importance in the crash database was the location (latitude and longitude) of crash occurrence. The location of the crash was used to overlay the crash data onto the transportation network of Hampton Roads region. Next, each crash was geocoded to one of the roadway segments (*i.e.,* spatial unit of analysis). Lastly, all crash occurrences on each roadway segment in the past year were aggregated to obtain the crash frequency that serves as the dependent variable of analysis.

4.2 Spatial Unit of Analysis

One of the first steps to crash frequency modeling is selecting the spatial unit of analysis, *i.e.* the geographical extent of region over which the expected crash frequency is modeled. The current study focusses on crash frequency along major interstates in the Hampton Roads region.

So, the empirical context implies that the interstates must be split into smaller segments that constitute the unit of analysis. However, this decision cannot be made arbitrarily because the availability of roadway inventory data and the homogeneity of resulting segments are critical to developing an accurate crash frequency model. So, several segment definitions were explored prior to choosing the spatial unit of analysis. For instance, the easiest and straightforward segment definition is uniform one-mile segments starting from the first mile marker of each interstate. However, such segmentation can result in non-homogenous segments, *i.e.* the roadway geometric characteristics and traffic conditions can vary considerably within each segment. For instance, a portion of the one mile stretch may correspond to the freeway portion and the remaining portion corresponds to ramp area. Another alternative was the publicly available Census Bureau's TIGER (Topologically Integrated Geographic Encoding and Referencing) database that divides each roadway into a contiguous stretch of several smaller segments. For instance, there were 72 unique TIGER Line segments along I-264 East in the study region. It is important to note that these segments are homogenous but not uniform. However, one of the limitations of using the TIGER segments was unavailability of extensive roadway inventory data. Barring a few important variables such as number of lanes and segment length, other key attributes such as shoulder and median presence were missing. The third alternative was using the segment definition in the VDOT's roadway inventory database that provided detailed information characterizing each segment. However, just as the TIGER segments, the VDOT segments were also not uniform. Based on the relative merits of the three segment definitions (uniform, TIGER, and VDOT), this study adopted the VDOT segment definition as the spatial unit of analysis. Figure 5 highlights the roadway segments used within the Hampton Roads Region in this study.

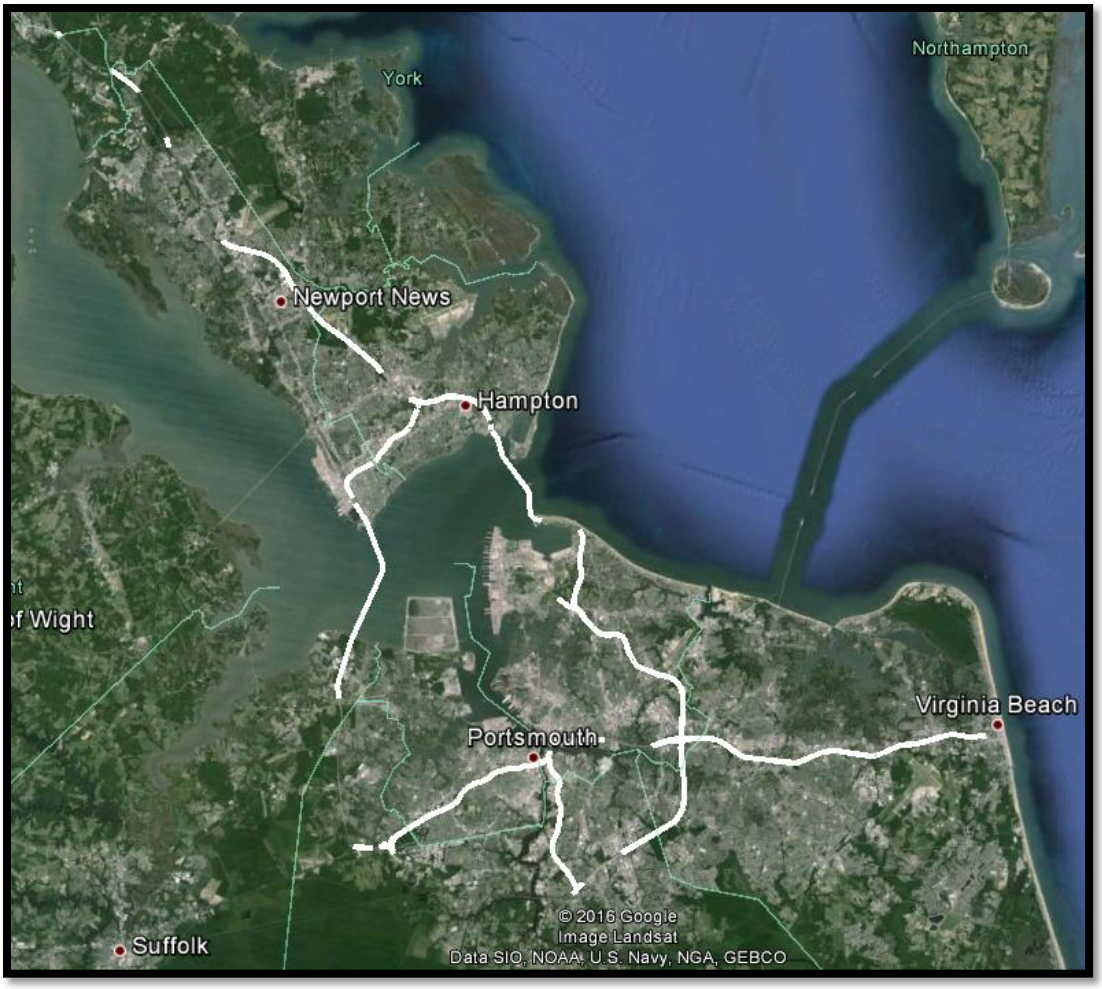

Figure 5. Roadway Segments Used in Analysis

4.3 Temporal Unit of Analysis

Weekend crashes were omitted due to travel patterns being inconsistent with other travel days. Also, a histogram of the crash data, seen in Figure 6, indicated that there was considerable over-representation of crashes during the peak period between 4:00 pm and 6:00 pm. Specifically, nearly 18% of all crashes in the past year occurred during the two hour PM peak period. This observation coupled with the constraint that it is not feasible to collect probe vehicle data using smartphones along all interstates during all hours of the day, the two hour time period

between 4 and 6 pm was chosen as the temporal unit of analysis. So, instead of using crash frequency in the past year along each roadway segment in the entire day, crashes that occurred during the two hour PM peak period were considered in the analysis. So, the dependent variable of analysis is crash frequency between 4 and 6 pm during weekdays in one year. Figure 7 displays the total number of trips recorded and crashes observed for each segment of roadway analyzed by this research.

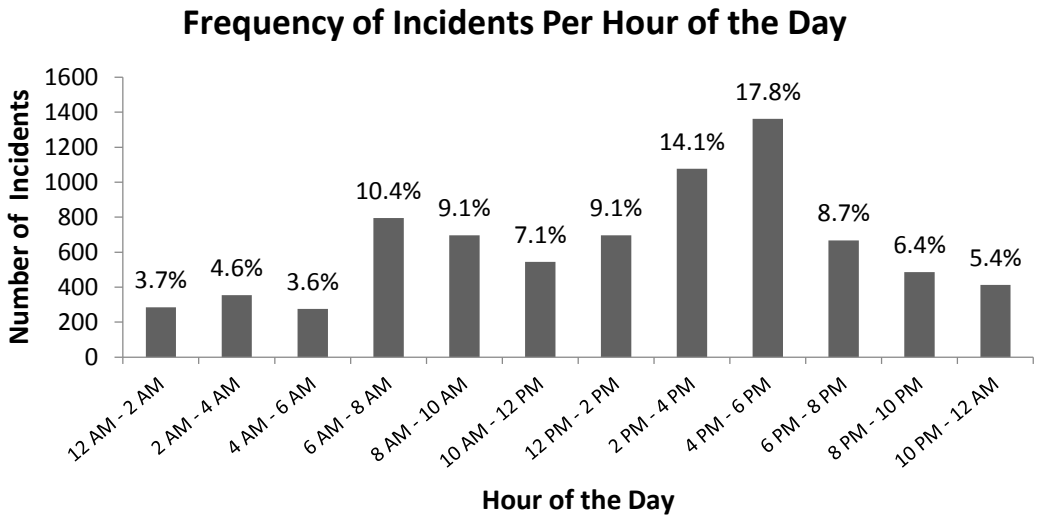

Figure 6. Total Incidents per Hour of the Day

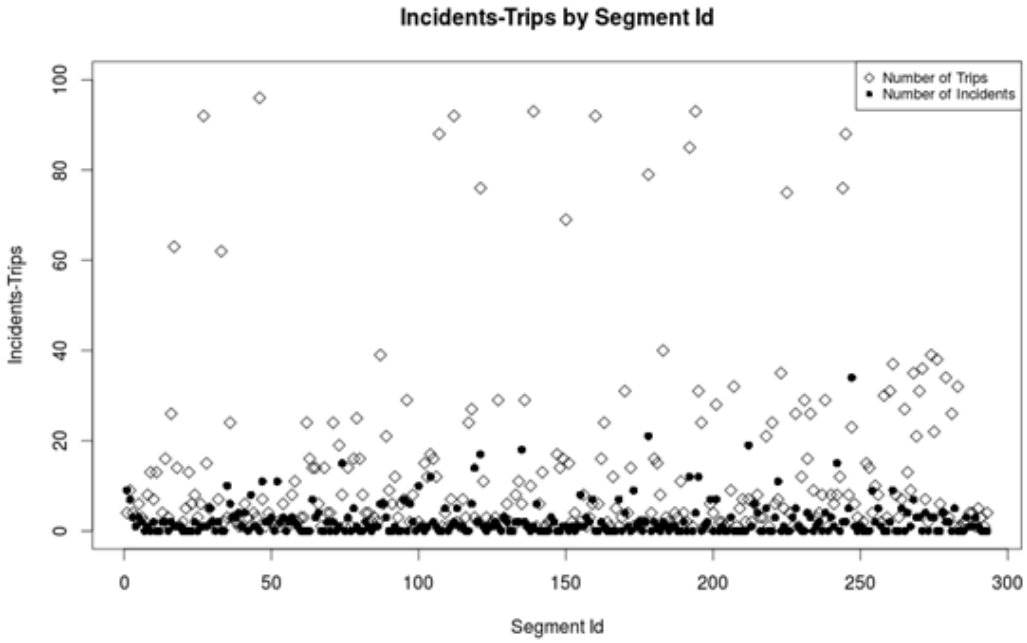

Figure 7. Total Number of Incidents and Trips by Segment Id

4.4 Supplementary Data

Several additional data sources were used to compile the explanatory variables in the crash frequency analysis. These data components can be sorted into three distinct categories: roadway inventory data, probe vehicle data, and exposure (*i.e.*, traffic volume).

The roadway inventory information was obtained from a VDOT maintained database that contains information regarding to the physical characteristics of the roadway. Some of the segments contained within the database contained incomplete or missing information. These segments were removed and there were a total of 293 unique roadway segments remaining to be used in this research. Figure 8 displays the frequency distribution for the number of crashes occurring in a single roadway segment during the two hour PM peak window.

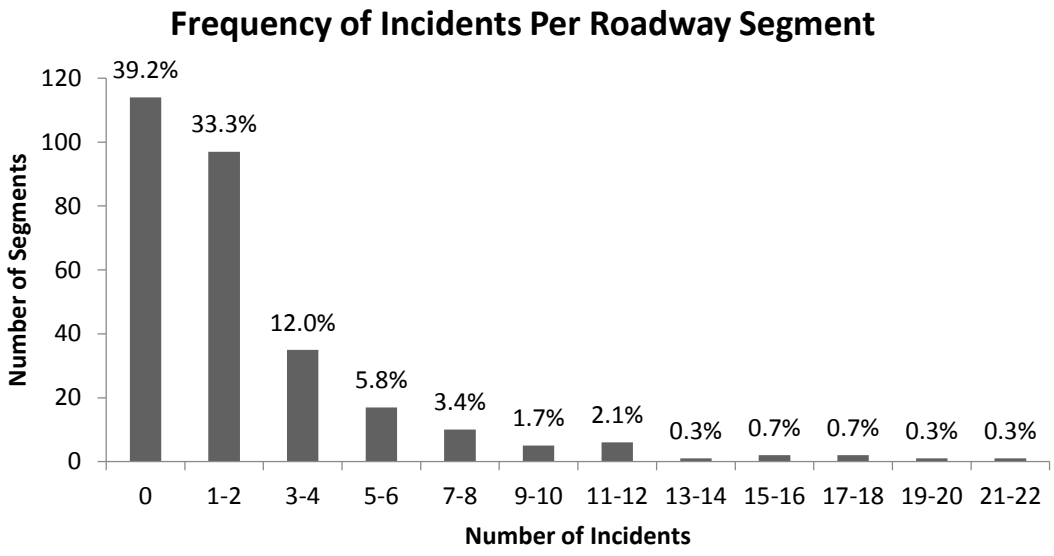

Figure 8. Frequency of Incidents per Roadway Segment

All of these segments fall on interstates within the Hampton Roads region of Virginia. The length of each segment was recorded to account for varying lengths between segments. The first piece of information pulled from this database is the number of lanes of the roadway segment. The number of lanes varied from one lane to five lanes. This variable was broken down into three separate categories: less than or equal to two lanes, three lanes, and greater than or equal to 4 lanes. The next variable used from this dataset was the surface type. This category was only broken down into two types within the roadway segments considered: plant mix and Portland cement concrete. The plant mix category is a typical asphalt roadway and the Portland cement concrete is a concrete surface. Surface width was taken from the database and broken into three categories: less than or equal to 24', 24' to 48', and greater than or equal to 48'. The presence of shoulder on both the right and left side of the roadway was also included in this database and recorded for analysis. If a shoulder is present, the width of the shoulder was also recorded and broken into three categories: Less than or equal to 8', 8'-12', and greater than or equal to 12'.

The database provided information as to whether or not the roadway segment was a high occupancy vehicle (HOV) lane or a regular lane. Along with HOV lanes, the database considered whether or not the lane was a reversible lane. These variables were considered in the model. Median presence was also considered and if there was a median, its type and size was considered. Types of median were split between grass median and a combination of positive barrier and curbed median for the analysis. The width of these medians was also considered. This category was broken into median widths which are less than 20', widths that are greater than or equal to 20' and less than or equal to 40', and widths that are greater than 40'. The final variable considered from the roadway inventory was the type of system the segment was contained in. This variable was broken into two categories: divided, full control of access, and a combination of roadways which were one-way, part of a one-way system, and two-way, non-divided roadways. Table 1 and Table 2 provide an overview of all of the previously mentioned roadway inventory explanatory variables, along with their frequency and percentage distributions used in the final dataset. Information regarding how these variables fit into the model can be found in the 'Results' section.

Table 1. Categorical Roadway Inventory Data

Number of Lanes	**Frequency**	**Percentage**
Less Than or Equal To 2	114	38.9%
3	92	31.4%
Greater Than or Equal to 4	87	29.7%
Total	293	100.0%
Surface Width		
Less Than or Equal To 24'	110	37.5%
24'-48'	96	32.8%
Greater Than or Equal to 48'	87	29.7%
Total	293	100.0%
Surface Type		
Plant Mix	140	47.8%
Portland Cement Concrete	153	52.2%
Total	293	100.0%
Presence of Right Shoulder		
Shoulder Present	144	49.1%
No Shoulder	149	50.9%
Total	293	100.0%
Right Shoulder Width		
Less Than or Equal To 8'	155	52.9%
8'-12'	137	46.8%
Greater Than or Equal to 12'	1	0.3%
Total	293	100.0%
Presence of Left Shoulder		
Shoulder Present	188	64.2%
No Shoulder	105	35.8%
Total	293	100.0%
Left Shoulder Width		
Less Than or Equal To 8'	143	48.8%
8'-12'	145	49.5%
Greater Than or Equal to 12'	5	1.7%
Total	293	100.0%
HOV Lane		
Lane is an HOV Lane	25	8.5%
Lane is not an HOV Lane	268	91.5%
Total	293	100.0%
Table Continued on Next Page		

Table 1. Continued

Reversible Lane	**Frequency**	**Percentage**
Lane is Reversible	2	0.7%
Lane is Non-Reversible	291	99.3%
Total	293	100.0%
Median Type		
Grass/Unprotected	87	29.7%
Positive Barrier or Curbed	94	32.1%
No Median	112	38.2%
Total	293	100.0%
Median Width Minimum		
Less than 20'	243	82.9%
Greater Than or Equal to 20' and Less Than or Equal to 40'	7	2.4%
Greater than 40'	43	14.7%
Total	293	100.0%
Facility Type		
One Way, Part of a One-Way System or Two-Way, Non-Divided	37	12.6%
Divided, Full Control of Access	256	87.4%
Total	293	100.0%

Table 2. Continuous Roadway Inventory Data

Continuous Variable	**Units**	**Mean**	**5th Percentile**	**95th Percentile**	**Standard Deviation**
Segment Length	Miles	0.44	0.09	1.20	0.44

The second major source of data was obtained through probe vehicles. Vehicles were equipped with cellular devices which were linked to on board diagnostic (OBD) devices through Bluetooth technology. The OBD device interfaces with the computer system within the vehicle itself. This device records information such as the velocity of the vehicle and the rotations per minute of the engine. The cellular device runs an Android application named GoGreen which has the capability of interfacing with the OBD device to record even more information through sensors located within the phone. The GoGreen application records the forces exerted on the cell phone by the moving vehicle by accessing the gyroscope inside of the phone. These vehicles

equipped with the data collection technology were driven on interstate roadways within Hampton Roads during the 4:00 p.m. to 6:00 p.m. time period. The probe vehicle data was collected by using the car following technique in which the probe vehicle drove at a speed very close the surrounding traffic in the right hand lane and passing slower traffic when feasible to mimic the "average" commuter. The 4 to 6 pm time period was previously selected based on the peak number of crashes occurring in this region during those hours. The GPS feature in the smartphone was also enabled to track vehicles as they drive along the interstates. The GPS coordinates were also used to map the probe vehicle onto the roadway segments that constitute the spatial unit of analysis. Several metrics were then calculated for each segment using the probe vehicle data to capture driving behavior.

First, the mean traffic speed for each segment was obtained by averaging the speed recordings for all trips contained within that single segment. This mean traffic speed was broken down into three distinct categories: less than 45 mph, greater than 45 mph and less than 60 mph, and greater than or equal to 60 mph. Next, speed data taken from the OBD device was used in order to calculate acceleration values for the model. The OBD device recorded speed values at a one second frequency. The difference between two consecutive velocity readings over a one second time period was considered the acceleration for that data point. This acceleration value was then converted to feet per second^2 for analysis. The acceleration data was also divided into two separate categories: accelerations and decelerations. Accelerations were taken as all positive acceleration recordings and decelerations were taken as all negative acceleration recordings. All positive acceleration values will be referred to as "Accelerations" and all negative accelerations will be referred to as "Decelerations" in the rest of the thesis. The minimum and maximum values of acceleration recorded were calculated in order to obtain information regarding the

extreme values observed within each roadway segment. Standard deviation was calculated across all accelerations and decelerations. Average acceleration values for were calculated among accelerations and decelerations separately as well as accelerations and decelerations combined and recorded into the dataset. To capture unique driving patterns, two additional metrics were calculated - the number of accelerations within a segment that were above median acceleration (NACC50) and the number of decelerations within a segment that were below median deceleration (NDEC50). These calculated values were also used to categorize segments based on their pattern of accelerations. If NACC50>0 and NDEC50=0, the segment is to be considered 'primarily accelerating'. Similarly, if NACC50=0 and NDEC50>0, the segment was considered 'primarily decelerating'. If NACC50>0 and NDEC50>0, then the segment was considered both accelerating and decelerating. Lastly, if NACC50=0 and NDEC50=0, the segment was considered steady flow due to the lack of higher end acceleration values. A similar method was taken to capture extreme driving patterns. The 5^{th} and 95^{th} percentile accelerations were calculated and if a segment had a deceleration recording below the 5^{th} percentile it was considered to have an extreme deceleration. If a segment had an acceleration recording above the 95^{th} percentile it was considered to have an extreme acceleration. If a segment had an acceleration above the 95^{th} percentile and a deceleration below the 5^{th} percentile then it was considered to have both extreme acceleration and deceleration present. The average speed was also calculated from probe vehicle recordings and included in this database. Table 3 displays the mean, 5^{th} percentile, 95^{th} percentile, and standard deviation for the continuous variables used in the final dataset.

Table 3. Continuous Metrics computed using Probe Vehicle Data

Continuous Variable	Units	Mean	5th Percentile	95th Percentile	Standard Deviation
Maximum Deceleration	ft/sec^2	-3.89	-9.25	-0.37	2.97
Maximum Acceleration	ft/sec^2	3.21	0.39	8.38	2.44
Ave. Accel. Across All Accel. And Decel.	ft/sec^2	0.01	-0.73	0.70	0.59
Average Deceleration	ft/sec^2	-0.01	-0.03	0.00	0.02
Average Acceleration	ft/sec^2	0.01	0.00	0.03	0.02
Standard Dev. Across Accel. And Decel.	ft/sec^2	1.15	0.32	2.13	0.60
Average Speed	mph	49.94	19.27	65.87	14.06

While estimating the count models, care was taken so that highly correlated continuous metrics (see Table 4) from the probe vehicle data are not simultaneously used in the model. As an example maximum deceleration was highly correlated with the standard deviation across accelerations and decelerations. This is expected because when there are higher maximum values then the standard deviation of a similar dataset is expected to be greater as well. Care was taken in model development to avoid including highly correlated metrics into the model.

Table 4. Correlation Matrix

	Max. Decel.	Max. Accel.	Average across all Accel. And Decel.	Average Decel.	Average Acel.	Standard Dev. Across Accel. And Decel.
Maximum Deceleration	1.000	-0.655	0.307	0.514	-0.458	-0.750
Maximum Acceleration	-0.655	1.000	0.154	-0.506	0.474	0.679
Average across all Accel. And Decel.	0.307	0.154	1.000	0.059	0.022	-0.152
Average Deceleration	0.514	-0.506	0.059	1.000	-0.960	-0.295
Average Acceleration	-0.458	0.474	0.022	-0.960	1.000	0.251
Standard Dev. Across Accel. And Decel.	-0.750	0.679	-0.152	-0.295	0.251	1.000

Table 5 displays the frequency and percentage distributions of the categorical variables used in the final dataset.

Table 5. Categorical Metrics computed using Probe Vehicle Data

Pattern of Accelerations	**Frequency**	**Percentage**
Primarily Accelerating	21	7.2%
Primarily Decelerating	29	9.9%
Both Accelerating and Decelerating	223	76.1%
Steady Flow	20	6.8%
Total	293	100.0%
Extreme Accelerations Present		
Yes	159	54.3%
No	134	45.7%
Total	293	100.0%
Table Continued on Next Page		

Table 5. Continued

Extreme Decelerations Present	**Frequency**	**Percentage**
Yes	166	56.7%
No	127	43.3%
Total	293	100.0%
Both Extreme Accel. And Decel. Present		
Yes	127	43.3%
No	166	56.7%
Total	293	100.0%
Average Speed		
Less Than 45 mph	57	19.5%
Greater Than 45 mph and Less Than 60 mph	117	39.9%
Greater Than or Equal to 60 mph	119	40.6%
Total	293	100.0%

The final source of data used in the study was exposure, *i.e.* the average traffic volume during the two hour peak period during weekdays in the past one year. This traffic volume was obtained by roadway sensors that are maintained by VDOT. These continuous count stations provide a means to control for traffic exposure levels along different roadway segments in the estimation dataset. Table 6 displays the mean, 5^{th} percentile, 95^{th} percentile, and standard deviation for the traffic exposure variable used in the final dataset.

Table 6. Traffic Volume

Continuous Variable	**Units**	**Mean**	**5th Percentile**	**95th Percentile**	**Standard Deviation**
Average Annual Weekday Peak Period Traffic	Vehicles	7862.22	2416.20	12576.00	8140.86

CHAPTER 5

RESULTS

All the models were estimated bottom-up by adding variables one at a time and checking statistical significance and intuitiveness at each step. As a general rule, 95% confidence rule was used for retaining parameters in the model. However, in some cases, parameter estimates with lower confidence level were also retained either if the corresponding result was intuitive or was deemed interesting to support future research. First, the Poisson model that is the most commonly used count model in the literature was estimated. Next, the Negative Binomial model that relaxes the equi-dispersion assumption of the Poisson model was estimated. Also, within each model, two versions were estimated. The first version corresponded to a model with only roadway inventory, speed, and volume variables that corresponds to typical crash frequency models in the literature. The second version corresponds to a model that also includes variables calculated using probe vehicles. This was done to demonstrate the improvement in the data fit provided by the probe-vehicle data unique to this study.

5.1 Poisson Regression Model

The first model created in an attempt to predict vehicle incidents followed a Poisson regression. Variables from the roadway inventory, traffic exposure, and probe vehicle data were all included in the model. Out of the initial roadway inventory variables tested, only 5 were deemed significant enough to remain within the model. Number of lanes was deemed insignificant for all categories. This most likely occurs because it has a strong correlation with the exposure variable which remained in the model. Surface width was also insignificant due to the same correlation with the exposure variable. The surface type of the roadway was also deemed insignificant and was removed. The presence of right shoulder was deemed significant

but, the width of the shoulder did not have a significant impact on the model. If a right shoulder is present within a roadway segment, the number of expected crashes is reduced. Similar to the right shoulder, the left shoulder also had an effect on predicting accident frequency. If a left shoulder is present then the expected number of crashes is reduced. The width of the shoulder did not have a significant impact on the model. The presence of an HOV lane was found to be insignificant and removed from the model. Reversible lane status was not significant in crash frequency predictions. The type of median had no significant impact on the model, but the median width did. It was found that there was not a significant difference between a median width less than 20' and a median width greater than or equal to 20' and less than or equal to 40' therefore, these two categories became a base case and the remaining category, greater than 40', remained in the model. If the segment had a median width minimum greater than 40' then it was deemed more accident prone than a roadway segment with a median width minimum less than or equal to 40'. The type of facility and the length of the segment were both deemed significant. If the facility type was considered 'One Way, Part of a One-Way System or Two-Way, Non-Divided' then there are more expected crashes than if the segment is on a facility considered 'Divided, Full Control of Access". This makes sense considering access control typically leads to less conflict points due to a decrease in access to the mainline. Lastly, as the length of the segment increases so does the expected number of crashes. This is intuitive due to there being more exposure to potential crashes in a longer roadway segment.

The next input to the Poisson model was the traffic exposure variable. Average annual weekday peak period traffic was deemed significant to the model and included. As the amount of traffic increases so do the expected number of crashes. This is to be expected considering that if there is more traffic on the roadway then the segment is exposed to more potential crashes.

Model results seen in Table 7 display parameter estimates if we were to consider only roadway inventory and exposure data, which most previous research has focused on.

Table 7. Initial Poisson Model Parameter Estimates

Parameter	Estimate	Z-Score
(Intercept)	1.530	2.401
Roadway Inventory Parameters		
ln(Segment Length)	0.666	14.440
Presence of Left Shoulder		
(Base: No Shoulder Present)		
Left Shoulder is Present	-1.170	-5.079
Presence of Right Shoulder		
(Base: No Shoulder Present)		
Right Shoulder is Present	-1.307	-5.518
Median Width Minimum		
(Base: Less Than or Equal To 40')		
Greater Than 40'	0.440	3.866
Facility Type		
(Base: Divided, Full Control of Access)		
One Way, Part of a One-Way System or Two-Way, Non-Divided	1.541	5.720
Exposure Parameter		
ln(Average Annual Weekday Peak Traffic)	0.122	1.938
Number of Cases	293.000	
Log Likelihood	-773.325	

Lastly, probe vehicle information was added to the model. Only four out of the initial probe vehicle variables calculated remained in the final Poisson model. Maximum acceleration and maximum declaration remained in the model. Both higher maximum accelerations and higher maximum decelerations lead to an increase in the predicted number of crashes. Although the parameter estimate on maximum decelerations is negative, it still leads to more predicted

crashes because the value for the variable ‘Max Deceleration’ will also always be negative. Average decelerations and average accelerations were both insignificant. Likewise, the average acceleration across decelerations and accelerations was also insignificant. These variables did not have much variation across roadway segments and were therefore not indicative of incidents. Standard deviation across decelerations and accelerations was also deemed insignificant and removed from the model. The number of decelerations above and below the overall median was not significant in this model. As a result, the categorical variables corresponding to the pattern of accelerations were also not significant. Extreme accelerations and extreme decelerations were both considered insignificant and removed from the model. An indicator for a segment having both extreme accelerations and extreme decelerations was also insignificant. Lastly, average traffic speeds were considered significant in the model. Segments which had average speeds less than 45 mph and segments which had an average speeds greater than or equal to 45 mph and less than 60 mph both lead to more expected crashes than segments which had average speed values which were greater than 60 mph.

The final parameter estimates, z-scores, number of cases, and log likelihood values can all be found in Table 8. In a later portion in this chapter we will discuss the impact that adding probe vehicle information had to the model instead of solely relying on roadway inventory data. There will also be a discussion as to how these results compare to the results of the negative binomial model.

Table 8. Final Poisson Model Parameter Estimates

Parameter	Estimate	Z-Score
(Intercept)	-0.387	-0.540
Roadway Inventory Parameters		
ln(Segment Length)	0.563	11.903
Presence of Left Shoulder		
(Base: No Shoulder Present)		
Left Shoulder is Present	-0.336	-1.752
Presence of Right Shoulder		
(Base: No Shoulder Present)		
Right Shoulder is Present	-0.600	-3.076
Median Width Minimum		
(Base: Less Than or Equal To 40')		
Greater Than 40'	0.301	2.579
Facility Type		
(Base: Divided, Full Control of Access)		
One Way, Part of a One-Way System or Two-Way, Non-Divided	0.502	2.134
Probe Vehicle Data Parameters		
Max Deceleration	-0.043	-2.338
Max Acceleration	0.131	6.129
Average Traffic Speed		
(Base: Greater Than or Equal to 60 mph)		
Less Than 45 mph	0.519	3.438
Greater Than or Equal to 45 mph and Less Than 60 mph	0.205	1.587
Exposure Parameter		
ln(Average Annual Weekday Peak Period Traffic)	0.133	1.816
Number of Cases	293.000	
Log Likelihood	-655.611	

5.2 Negative Binomial Regression Model

Once the Poisson model was completed, a negative binomial model was estimated in an effort to obtain a better fit given that preliminary descriptive analysis showed over-dispersion in the crash data. This model was able to reduce the number of roadway inventory variables from five to one. The only remaining roadway inventory parameter that was deemed significant in this

model was the length of the segment. As expected, a greater segment length leads to more predicted crashes in this initial model. This makes sense considering there is more lane miles for a crash to occur over. This model also included the exposure parameter of average annual weekday peak traffic. As the traffic volumes increase so do the expected number of crashes. This intuitive behavior can be observed across all models. Table 9 displays the model estimated with only roadway inventory and exposure variables.

Table 9. Initial Negative Binomial Model Parameter Estimates

Parameter	**Estimate**	**Z-Score**
(Intercept)	0.727	0.626
Roadway Inventory Parameters		
ln(Segment Length)	0.687	6.823
Exposure Parameter		
ln(Average Annual Weekday Peak Traffic)	0.089	0.678
Dispersion Parameter	0.711	7.705
Number of Cases	293.000	
Log Likelihood	-564.957	

After the initial model was estimated, probe vehicle data was added to the model. This model reduced the number of probe vehicle variables from four to three when compared to the Poisson model. Maximum acceleration was included in the model and a higher maximum acceleration leads to a higher predicted crash frequency. An indicator variable for segments which have both extreme accelerations and extreme decelerations present was also found to be significant to this model. If a segment has these types of accelerations present then it is expected to have more crashes occurring. Lastly, average traffic speed was included in the model. If the

average traffic speed is less than 45 mph then the segment is expected to have a higher number of crashes than if the segment has an average speed which is greater than or equal to 45 mph. Table 10 displays the final parameter estimates, z-scores, log likelihood and number of cases for the final negative binomial model.

Table 10. Final Negative Binomial Model Parameter Estimates

Parameter	**Estimate**	**Z-Score**
(Intercept)	0.093	0.085
Roadway Inventory Parameters		
ln(Segment Length)	0.582	6.403
Probe Vehicle Data Parameters		
Max Acceleration	0.100	2.232
Extreme Accelerations and Decelerations Present		
(Base: No)		
Yes	0.401	1.907
Average Traffic Speed		
(Base: Greater Than or Equal to 45 mph)		
Less Than 45 mph	0.329	1.707
Exposure Parameter		
ln(Average Annual Weekday Peak Period Traffic)	0.063	0.515
Dispersion Parameter	1.021	6.630
Number of Cases	293.000	
Log Likelihood	-540.266	

5.3 Statistical Fit Comparisons

Statistical fit comparisons were conducted in order to compare the effectiveness of the different models which were created. The first goal was to ensure that adding probe vehicle data to the analysis actually improved the overall fit of the data. A log likelihood ratio test was implemented for this purpose given that the models with and without probe vehicle variables are nested versions of each other. Two times the difference between the Log Likelihood of the

Poisson model with only roadway geometry variables to the Log Likelihood of the full Poisson model, including probe vehicle data, was computed to be 235. This value was then compared with the critical chi-squared value corresponding to the additional degrees of freedom in the un-restricted model. There were four degrees of freedom between the two models and the corresponding critical chi squared value is 9.488. Considering 235 is greater than 9.488, adding the probe vehicle data improved the model significantly. Likewise, when comparing the two versions of the negative binomial model, the Log Likelihood ratio was calculated to be 49 and the critical chi squared value corresponding to two degrees of freedom is 5.991. Adding probe vehicle data to the negative binomial model also improved the model significantly considering 49 is greater than 5.991.

The next goal was to determine whether or not switching from a Poisson model to a negative binomial model improved the results. Considering that the negative binomial model and the Poisson model are not nested versions of each other, using the standard log likelihood ratio test is not acceptable. A Bayesian Information Criterion (BIC) test is typically used when two models are non-nested versions of each other. A model with lower BIC value is preferred over the other model. The BIC values were computed for the final Poisson and negative binomial models as $-2 \times Log - Likelihood + k \times LN(N)$, where k is the number of model parameters and N is the size of the dataset. The BIC value for the Poisson model was calculated to be 1368 while the BIC value for the negative binomial model was calculated to be 1109. Considering the BIC value for the negative binomial model is less than the BIC value for the Poisson model, it is to be considered the more appropriate model for analysis.

5.4 Elasticity Effects

While the parameter estimates discussed in the earlier section indicate the directionality of different factors, it is difficult to understand the magnitude of the effect by just looking at the parameters. So, elasticity effects of each variable were calculated in an effort to determine the effect each variable has on the final dataset. This will allow for the interpretation of how a percent increase in the parameter estimate will impact the expected number of crashes. Considering the model has both continuous and categorical variables, a slightly different approach had to be taken with both. For continuous variables, we first calculated the expected number of crashes for each roadway segment in the dataset. Then, we calculated expected crash frequency for each roadway segment after increasing the variable for which the elasticity was being computed by 100%. Next, percentage change between final and initial crash frequency was computed for each roadway segment. Lastly, this percentage change was averaged across all roadway segments to obtain the average elasticity effect of the corresponding variable. The resulting percentage value represents how a 100% increase in the corresponding variable will impact the expected number of crashes (everything else being the same). For example, if the maximum deceleration variable has an elasticity of 19%, a 100% increase in the maximum deceleration variable will result in a 19% increase in the expected number of crashes.

Elasticity effects for categorical variables (in which dummy variables were used) were computed using a slightly different approach. First, expected crash frequency was calculated assuming that all the indicator variables corresponding to the categorical variable assume a value of 0. Next, expected crash frequency was recomputed after changing the indicator variable corresponding to the category being considered to 1. The percentage difference between the two expected crash frequencies was reported as the elasticity effect of the corresponding category.

For instance, when considering the 'Presence of Left Shoulder' categorical variable, the base case is 'No Shoulder Present' and there is a dummy variable- 'Left Shoulder is Present'. In order to calculate the elasticity for the 'Left Shoulder is Present' category, we first assume that this indicator variable takes a value of zero and calculate the expected number of crashes. We then do the calculation again but this time after changing the indicator variable for 'Left Shoulder is Present' to 1. We then calculate the percentage difference between the two expected crash frequency values for each roadway segment and average the percentage change across all segments to obtain an elasticity value. The elasticity effect of 'Left Shoulder is Present' is -29% in the Poisson model. The interpretation of this value is that, on average, roadway segments which have a left shoulder present have 29% fewer crashes than roadway segments which do not have a left shoulder, everything else being same. This same logic can be applied to all other categorical variables and can be interpreted similarly.

Elasticity effects for the final Poisson model were calculated first. All interpretations assume that all other variables remain the same and only the targeted variable is changing. It was determined that a 100% increase in the segment length would cause a 48% increase in the expected number of crashes. When considering the presence of a left shoulder, a segment with a left shoulder present has a 29% decrease in expected crash frequency compared to a segment without a left shoulder. Next, the presence of a right shoulder was interpreted. A segment which has a right shoulder should expect a decrease in crash frequency of 45% when compared to a segment which does not have a right shoulder. If the minimum median width is greater than 40' then there is a 35% increase in expected crashes than when the median width minimum is less than or equal to 40'. If the segment is on a facility which is considered a 'One Way, Part of a One-Way System or Two-Way, Non-Divided' then there is an expected 65% increase in crashes

than if the segment is on a facility considered ‘Divided, Full Control of Access”. A 100% increase in maximum deceleration results in 19% increase in expected crash frequency while a 100% increase in maximum acceleration results in a 61% increase in expected crash frequency. If the average traffic speed for a segment is less than 45 mph then there is an expected increase in crash frequency of 68% when compared to traffic which has an average speed of greater than or equal to 60 mph. Likewise, if the average traffic speed for a segment is greater than or equal to 45 mph and less than 60 mph, there is an expected increase in crash frequency of 23% when compared to average speeds which are greater than or equal to 60 mph. This is intuitive considering slower traffic speeds typically suggest less congestion. As congestion increases so does the likelihood of a crash occurring. Lastly, if the average annual weekday peak period traffic increases by 100% then there is an expected increase in crash frequency of 10%. These elasticity values can be observed in Table 11 for the Poisson model.

Table 11. Final Poisson Model Elasticity Effects

Parameter	Elasticity
Roadway Inventory Parameters	
ln(Segment Length)	48%
Presence of Left Shoulder	
(Base: No Shoulder Present)	
Left Shoulder is Present	-29%
Presence of Right Shoulder	
(Base: No Shoulder Present)	
Right Shoulder is Present	-45%
Median Width Minimum	
(Base: Less Than or Equal To 40')	
Greater Than 40'	35%
Facility Type	
(Base: Divided, Full Control of Access)	
One Way, Part of a One-Way System or Two-Way, Non-Divided	65%
Probe Vehicle Data Parameters	
Max Deceleration	19%
Max Acceleration	61%
Average Traffic Speed	
(Base: Greater Than or Equal to 60 mph)	
Less Than 45 mph	68%
Greater Than or Equal to 45 mph and Less Than 60 mph	23%
Exposure Parameter	
ln(Average Annual Weekday Peak Period Traffic)	10%

Elasticity effects for the negative binomial model were calculated in the same manner as the Poisson model. A 100% increase in segment length results in a 50% expected increase in crash. A 100% increase in maximum acceleration results in a 42% increase in expected crash frequency. If extreme accelerations and extreme decelerations are observed within a segment then there is an expected crash frequency increase of 49% when compared to segments which do not have these extreme behaviors present. If the average traffic speed for a segment is less than 45 mph then there is an expected increase in crash frequency of 38% when compared to traffic

which has an average speed of greater than or equal to 45 mph. Again this is intuitive considering lower speeds typically relates to more congestion which causes a higher crash frequency. Lastly, if the average annual weekday peak period traffic increases by 100% then there is an expected increase in crash frequency of 4%. These elasticity values can be observed in Table 12 for the negative binomial model.

Table 12. Final Negative Binomial Model Elasticity Effects

Parameter	**Elasticity**
Roadway Inventory Parameters	
ln(Segment Length)	50%
Probe Vehicle Data Parameters	
Max Acceleration	42%
Extreme Accelerations and Decelerations Present	
(Base: No)	
Yes	49%
Average Traffic Speed	
(Base: Greater Than or Equal to 45 mph)	
Less Than 45 mph	39%
Exposure Parameter	
ln(Average Annual Weekday Peak Period Traffic)	4%

CHAPTER 6

CONCLUSIONS

Crash frequency modelling is a complex task with intricacies lying within the specific variables used and the type of model implemented. The majority of previous research focused primarily on roadway inventory data as the primary source of explanatory variables along with limited exposure variables such as traffic volumes. This research sought to expand on these previous models by using probe vehicle data in conjunction with roadway inventory and exposure data. This research took advantage of newer technologies to capture probe vehicle data and provide a more robust input into the crash frequency model. The results of this study included the following key findings:

1. Adding probe vehicle data to the model improved model results significantly. This improvement was observed when implementing both a Poisson model and a negative binomial model. Probe vehicle information provided data from vehicles which were actually travelling down the roadway. Microscopic traffic measures that characterize driving patterns based on acceleration and deceleration profiles were calculated using the accelerometer sensors in smartphones. Combining this information with existing roadway inventory data provided the model with a new category of information that improved the statistical data fit considerably.
2. Crash frequency modelling typically involves data which is over-dispersed in nature. This research implemented a negative binomial model in order to properly assess over-dispersion. The negative binomial model was considered a significant improvement to the initial Poisson model. The results also indicate considerable differences in the parameter estimates of the Poisson and negative binomial models. To be specific, the results

indicate that in the absence of the additional degree of freedom provided by the dispersion parameter in the negative binomial model, the Poisson model seems to compensate by over-estimating the significance of several explanatory variables. In comparison, the final model specification of the NB model was much more parsimonious (*i.e.*, fewer parameters) with considerably better log-likelihood.

Using probe vehicle data in conjunction with implementing a negative binomial model improved upon previous research which relied on roadway inventory and exposure data. While this study was successful in estimating an effective crash frequency model, there were some opportunities missed for potential research extension due to time constraints. The following are major potential improvements to this research:

1. This study calculated and considered over 50 potential explanatory variables for predicting incidents. However, there were still a few more critical variables which should be calculated to provide an even more complete model. Future model extension should consider the following variables in addition to those already tested:
    a. The distance to and from the nearest on and off ramp – Many incidents occur due to merging traffic at on and off ramps. These calculated variables will help spatially connect the expected influx of incidents at these locations.
    b. The driving patterns along a road are related to the driving patterns along the upstream and downstream segments. This spatial dependency in driving patterns (and thus crash frequencies) can be captured using spatial proximity weight matrix. This weight matrix is a square matrix of dimension equal to the number of roadway segments in the dataset. Each cell element is inversely related to the distance between the two segments with the underlying idea being that farther the

segment lower would be the influence. This weight matrix would then be multiplied with the vector of metrics calculated using probe vehicles. The resulting new vector of spatially weighed variables will serve as additional explanatory variables. The parameter estimates on these spatially weighted or lagged explanatory variables capture the spatial dependency among roadway segments in the region.

c. One of the interesting results in this study was that traffic exposure had a positive effect on crash frequency but was not significant in the negative binomial model. It was, however, significant in the Poisson model. Currently, the traffic volume data was obtained using VDOT sensors that were distributed all over the interstates in the Hampton Roads region. However, these sensors are dense enough to cover all the roadway segments in our study. Specifically, there were fewer sensors than segments and we had to use some weighing techniques to estimate the traffic volumes for some of the roadway segments. It is possible that this approximation has biased the effect of traffic exposure in the model. Future studies may uncover significant traffic exposure effect by using better data sources.

2. The negative binomial model was an improvement on the Poisson model but, there is room for even further improvement. Considering how rare incidents are, there are many segments which had a total of zero incidents over the entire yearlong study period. In fact, 39.2 % of the roadway segments had zero crashes. So, there is considerable over-representation of zeroes in the crash dataset. Unfortunately, negative binomial models may not always capture this excess zeroes problem (Chin and Quddus, 2003 and Lord *et*

al., 2005). In many cases, zero inflation is the reason for over-dispersion in the dataset. In such cases, the negative binomial model is adequate for modeling crash frequency. In other cases, a zero inflated negative binomial model is needed to account for both over-dispersion and extra zeros in the dataset. There are also some recent methodological advancements such as the Generalized Ordered Response Probit (GORP) and the GEV count models that were provide better flexibility than zero-inflated and hurdle models (Castro *et al*., 2012 and Paleti, 2016). Future extensions of this study may compare the performance of standard count models against these more advanced models.

3. All the models developed in this study are fixed-parameter models. So, they do not capture unobserved heterogeneity in the parameter estimates. However, in reality, there can be several unobserved factors that can moderate the effect of all factors that influence crash occurrences. For instance, the effect of speed on one roadway segment can be very different from the effect of speed on another roadway segment. This unobserved heterogeneity in parameter effects can be captured using random effects or mixing models that assume a distribution (ex: normal distribution) and estimate both the mean and the standard deviation of the parameter estimate (Lord and Mannering, 2010 and Anastasopoulos and Mannering, 2009). Future extensions must evaluate these mixing models.
4. Probe vehicle data collection is currently underway for additional segments that were not included in this study. It would useful to undertake a validation exercise on a separate dataset (not used for model estimation) by predicting the crash frequency along these segments and comparing with the observed crash frequency distribution. Such a

validation exercise will serve as an additional validation of the result that probe vehicle data improves the model accuracy.

BIBLIOGRAPHY

af Wåhlberg, A. E. (2004). The stability of driver acceleration behavior, and a replication of its relation to bus accidents. *Accident Analysis & Prevention, 36*(1), 83-92.

Ahmed, M., Huang, H., Abdel-Aty, M., & Guevara, B. (2011). Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention, 43*(4), 1581-1589.

Amoros, E., Martin, J. L., & Laumon, B. (2006). Under-reporting of road crash casualties in France. *Accid Anal Prev, 38*(4), 627-635.

Anastasopoulos, P. C., & Mannering, F. L. (2009). A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention, 41*(1), 153-159.

Blincoe, L., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015). *The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised)*.

Castro, M., Paleti, R., & Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation research part B: methodological*, *46*(1), 253-272.

Chin, H. C., & Quddus, M. A. (2003). Modeling count data with excess zeroes an empirical application to traffic accidents. *Sociological methods & research*, *32*(1), 90-116.

DMV (2014). 2014 Virginia Traffic Crash Facts, Virginia Highway Safety Office, Department of Motor Vehicles, Commonwealth of Virginia.

Eustace, D., Aylo, A., & Mergia, W. Y. (2015). Crash frequency analysis of left-side merging and diverging areas on urban freeway segments – A case study of I-75 through downtown Dayton, Ohio. *Transportation Research Part C: Emerging Technologies, 50*, 78-85.

FHWA (2015). FHWA Forecasts of Vehicle Miles Traveled (VMT): May 2015, Office of Highway Policy Information, Federal Highway Administration

Gettman, D., & Head, L. (2003). Surrogate Safety Measures from Traffic Simulation Models. *Transportation Research Record: Journal of the Transportation Research Board, 1840*, 104-115.

Hinde, J., & Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics & Data Analysis, 27*(2), 151-170.

Jun, J. (2006). *Potential Crash Exposure Measures Based on GPS-Observed Driving Behavior Activity Metrics.* The Georgia Institute of Technology.

Kim, K., Nitz, L., Richardson, J., & Li, L. (1995). Personal and behavioral predictors of automobile crash and injury severity. *Accident Analysis & Prevention, 27*(4), 469-481.

Li, Z., Wang, W., Liu, P., Bigham, J. M., & Ragland, D. R. (2013). Using Geographically Weighted Poisson Regression for county-level crash modeling in California. *Safety Science, 58*, 89-97.

Lord, D., Geedipally, S. R., & Guikema, S. D. (2010). Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Underdispersion. *Risk Analysis, 30*(8), 1268-1276.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291-305.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention, 37*(1), 35-46.

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research, 1*, 1-22.

Mekker, M. M., Remias, S. M., McNamara, M. L., & Bullock, D. M. (2015). *Characterizing Interstate Crash Rates Based on Traffic Congestion Using Probe Vehicle Data*. Paper presented at the Transportation Research Board 95th Annual Meeting.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention, 26*(4), 471-482.

Narayanamoorthy, S., Paleti, R., & Bhat, C. R. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B: Methodological, 55*, 245-264.

NHTSA (2014). Traffic Safety Facts 2012 Data, National Highway Traffic Safety Administration

NHTSA (2015). Traffic Safety Facts Crash-Stats, National Highway Traffic Safety Administration

Ogle, J. H. (2005). *Quantitative assessment of driver speeding behavior using instrumented vehicles.*

Paleti, R. (2016). Generalized Extreme Value models for count data: Application to worker telecommuting frequency choices. *Transportation Research Part B: Methodological, 83*, 104-120.

Qin, X., Ivan, J., Ravishanker, N., & Liu, J. (2005). Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling. *Journal of Transportation Engineering, 131*(5), 345-351.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention, 29*(6), 829-837.

Shinar, D., Treat, J. R., & McDonald, S. T. (1983). The validity of police reported accident data. *Accident Analysis & Prevention, 15*(3), 175-191.

Yamamoto, T., Hashiji, J., & Shankar, V. N. (2008). Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accid Anal Prev, 40*(4), 1320-1329.

Zhen, C., & Qiang, G. (2014). Mobile Sensor Data Collecting System Based on Smart Phone. In Q. Zu, M. Vargas-Vera & B. Hu (Eds.), *Pervasive Computing and the Networked World: Joint International Conference, ICPCA/SWS 2013, Vina del Mar, Chile, December 5-7, 2013. Revised Selected Papers* (pp. 8-14). Cham: Springer International Publishing.

VITA

Mr. Kenneth Wynne is a graduate student in the Department of Civil and Environmental Engineering at Old Dominion University. He is pursuing a Master of Science Degree in Civil Engineering with a focus in the Transportation field. He currently holds a Bachelor of Science Degree in Civil and Environmental Engineering from Virginia Polytechnic Institute and State University (May 2013). He began his professional career as an Engineering Scholar at the Virginia Department of Transportation (VDOT) in the summer of 2011. Upon completing his undergraduate studies he began working as a Transportation Engineer I at VDOT. In the winter of 2015 he took on the role of Project Manager at VDOT's Hampton Roads Project Management Office.

Kenneth Wynne

Department of Civil and Environmental Engineering

Frank Batten College of Engineering and Technology

Old Dominion University, Norfolk, VA 23529